The University of Manchester

Estimating mutual information in under-reported variables

K. Sechidis[1], M. Sperrin[2], E. Petherick[3], G Brown[1]

[1]School of Computer Science, University of Manchester (UK)
[2]Centre for Health Informatics, Institute of Population Health, University of Manchester (UK)
[3]School of Sport, Exercise & Health Sciences, Loughborough University (UK)

Main idea

Estimate the correlation between

Maternal Smoking and Low birthweight

Collect accurate measurements: **expensive/privacy**

Self-reported data, such as Born In Bradford project

Under-reporting (UR) bias

Non-smokers always tell the truth,

while smokers may lie

Main idea

Estimate mutual information between

Y: low birth weight of an infant $Y=\{0,1\}$

X: maternal smoking $X=\{0,1\}$

Population value
$I(X;Y)=0.12$ nats

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \widehat{p}(x,y) \ln \frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)}$$

Point estimate
$I(X;Y)=0.15$ nats

$$SE\left[\widehat{I}(X;Y)\right] = \frac{\sigma_{MI}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 - I(X;Y)^2 \right)^{\frac{1}{2}}$$

Interval estimate
95% Confidence Interval
$I(X;Y)\in[0.10-0.20]$ nats

Main idea

Estimate mutual information between

Y: low birth weight of an infant $Y=\{0,1\}$

X: maternal smoking $X=\{0,1\}$

…. **but** it is more convenient to collect self reported data:

X : the mother reported smoking or not $X=\{0,1\}$

… *I(X;Y)* ?

Misclassification bias problem

UR can be seen as a special case of **misclassification bias**

Epidemiology: Corrections for the odds-ratio and relative risk

using knowledge over specificities/sensitivities

Specificity: Pr ($X=0 \mid \mathcal{X}=0,$) **= 1** **Under reported**

Sensitivity: Pr ($X=1 \mid \mathcal{X}=1,$) **< 1** **Scenario**

Our work: Correction for mutual information

X: Reported smoking / $\mathcal{X}$: Actual smoking

Biases can be seen as missing data problems

UR can be seen as a special case of **positive-unlabelled** (PU)

a restricted semi-supervised binary problem

- Labelled set: only positive examples (Y=1)

  cases reported smoking

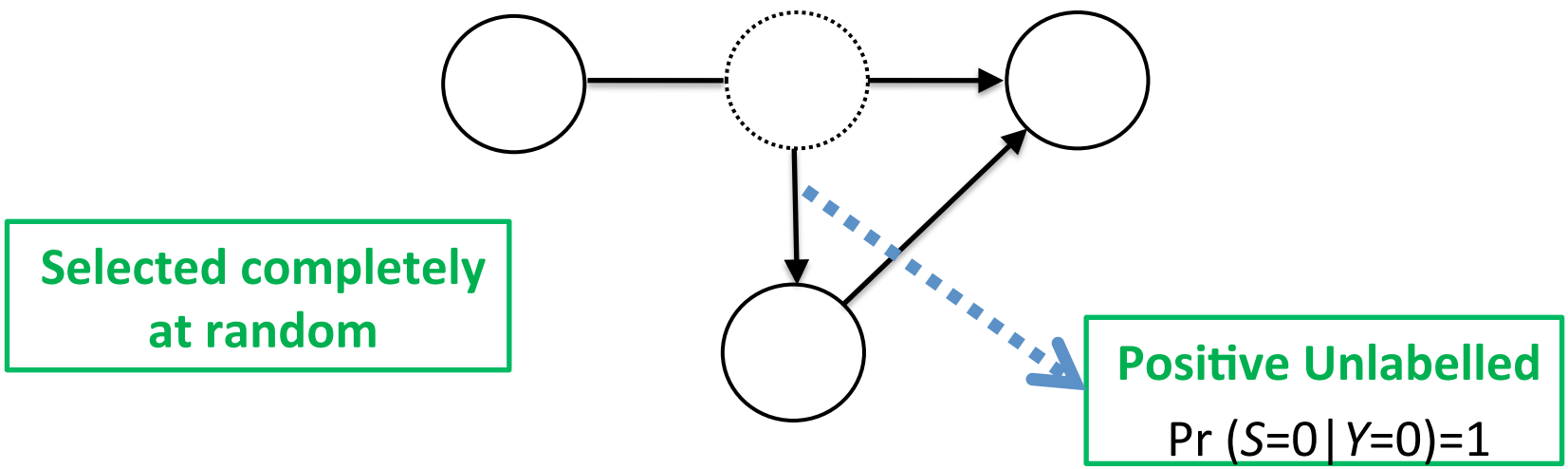
- Unlabelled set: either positive/negative (Y=0 or Y=1)

  cases reported non-smoking

using knowledge over prior P($Y=1$)

Missingness graphs for PU data

Missingness graphs (Pearl et al. 2013-2015)



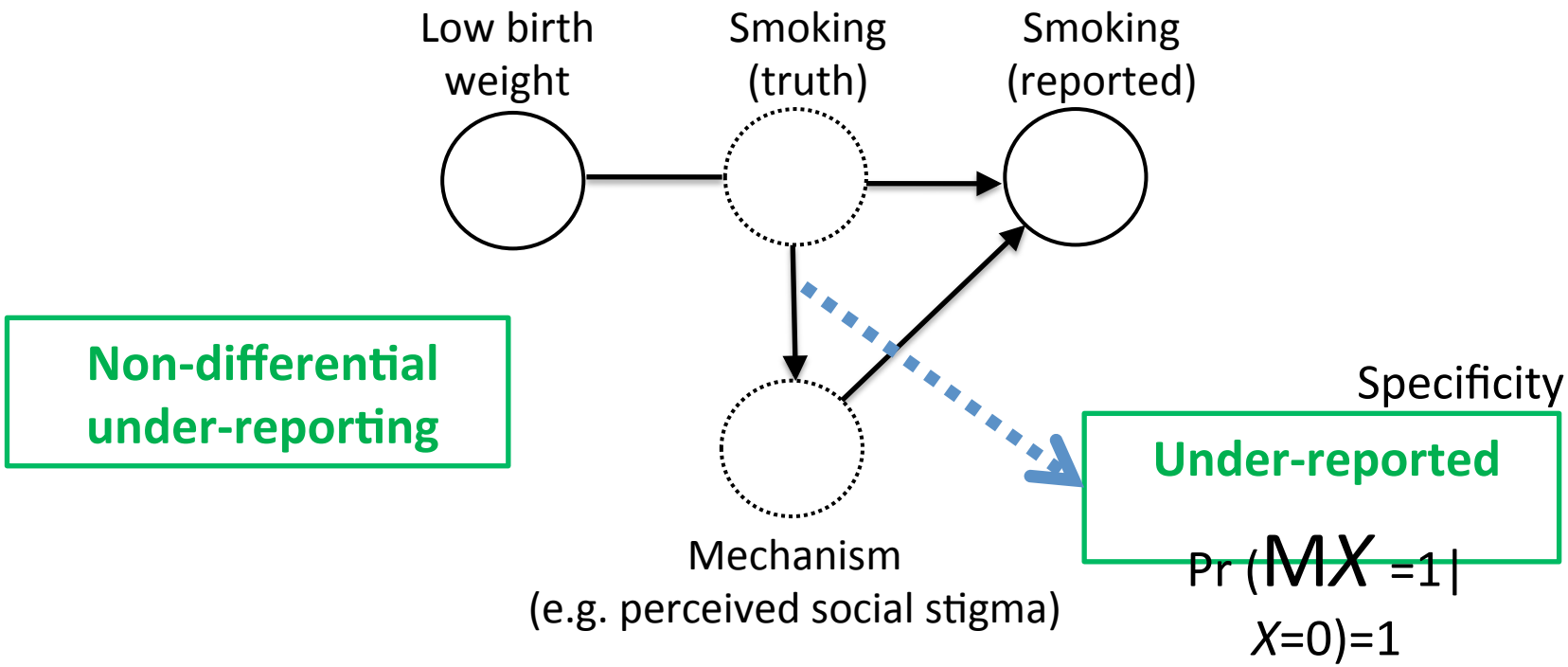
S: labelling mechanism $S=\{0,1\}$:1 labelled

0 unlabelled

Y:observed variable $\{0,1,\mathrm{m}\}$

Graph representation for UR data

Misclassification graphs

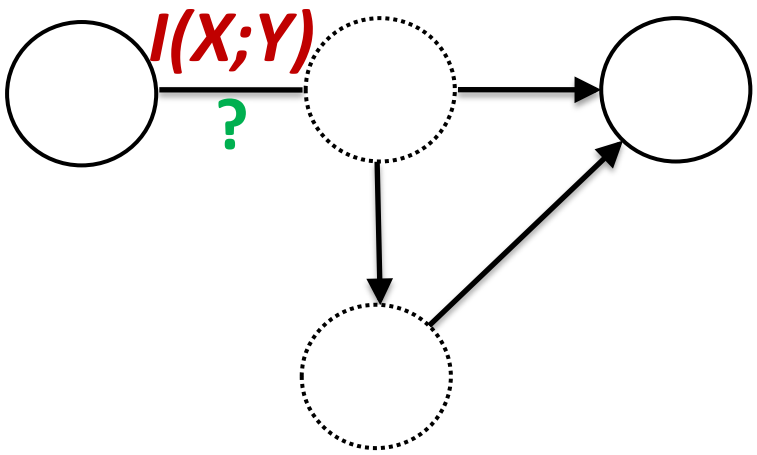


MX: misclassification mechanism $MX=\{0,1\}$, 1 correctly reported

0 misclassified

X:observed variable $\{0,1\}$

Mutual information in UR scenarios



No correction

$I(X;Y)$

Ideal

$I(X;Y)$

$$I(X;Y) < I(X;Y)$$

- ***Correct X:*** Use this model to impute values for the possible misclassified examples: women that reported non-smoking.
- **Correct MI directly:** Derive a corrected estimator that takes into account the under-reporting.

Correcting Mutual Information for UR

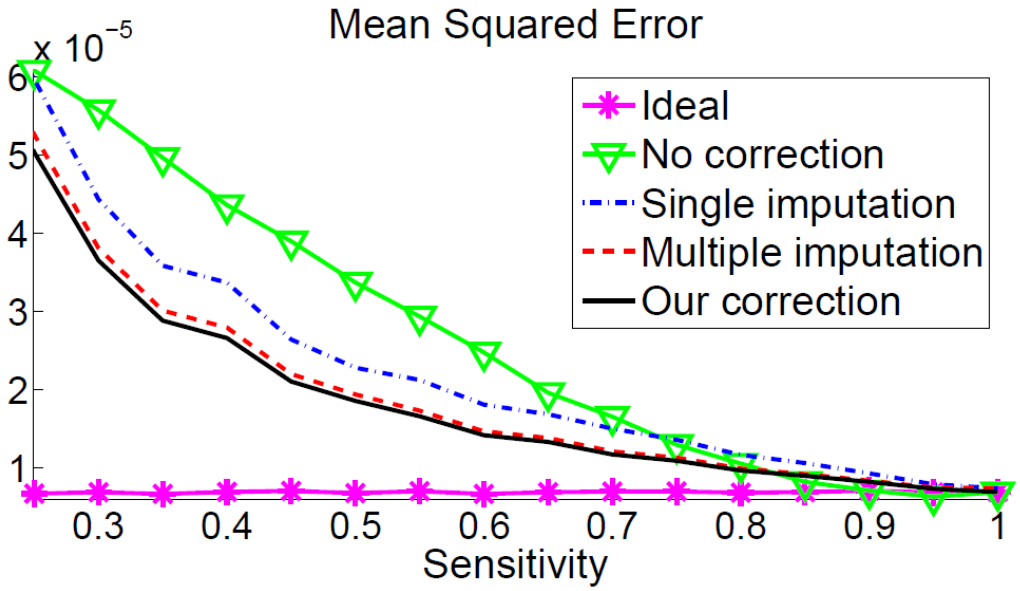
$$\widehat{I}_\gamma(\widetilde{X};Y) = \sum_{y\in\mathcal{Y}} \left(\gamma\, \widehat{p}(y|\widetilde{x}=1) \ln \frac{\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)} + (\widehat{p}(y) - \gamma\widehat{p}(y|\widetilde{x}=1)) \ln \frac{\widehat{p}(y) - \gamma\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)\,(1-\gamma)} \right)$$

This estimator is consistent when we have perfect knowledge over the prior: $\gamma = p(x=1)$

$$I_{\downarrow\gamma}(X;Y) = I(X;Y)$$

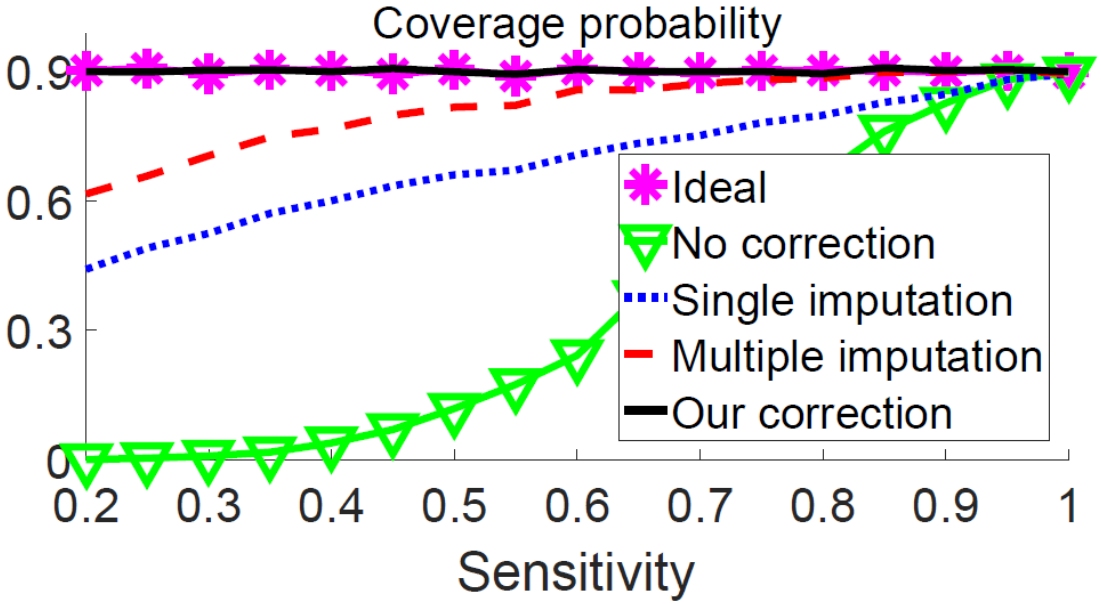
Known asymptotic distribution

Perfect Prior Knowledge

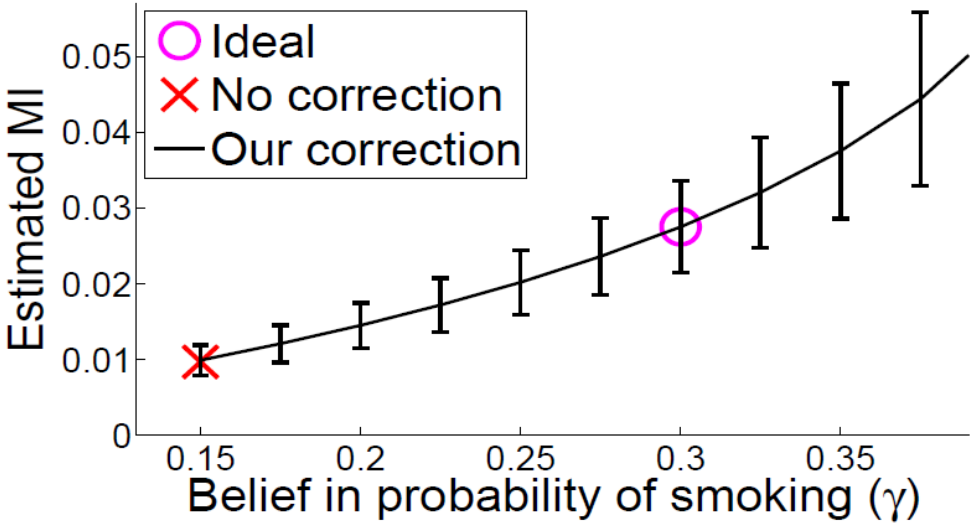


Comparison in terms of the **Mean Squared Error**

Comparison in terms of the **coverage of the 90% confidence interval**

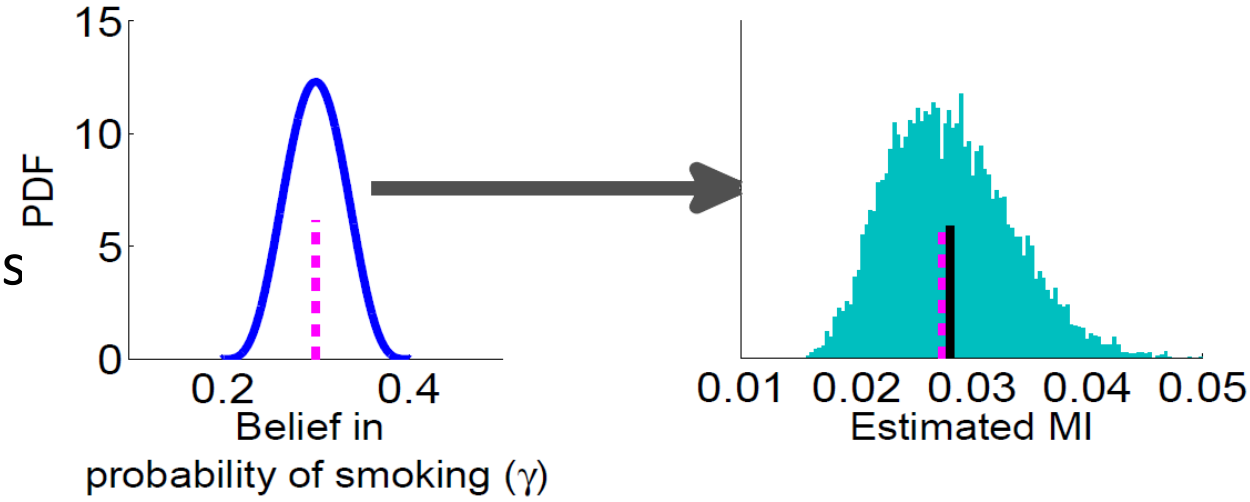


Uncertain Prior Knowledge

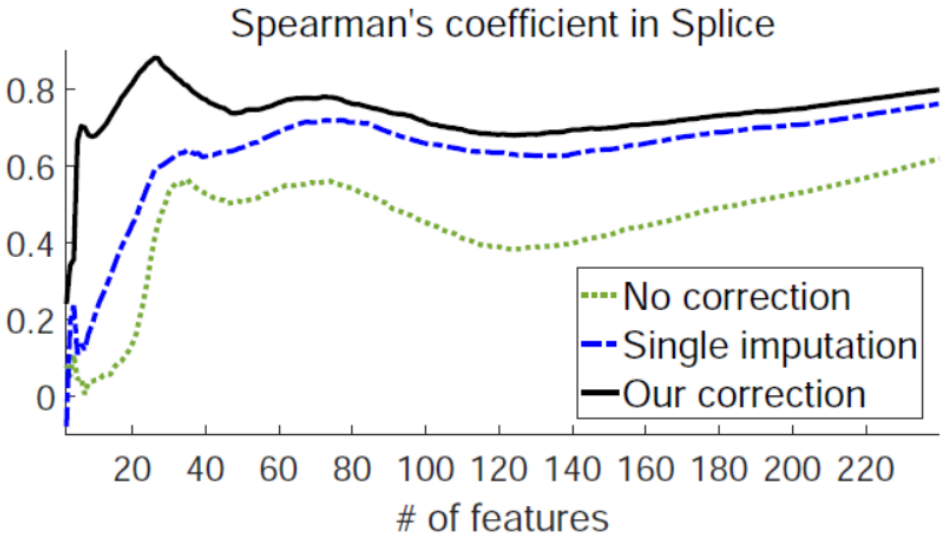


Sensitivity analysis

Simulation based analysis

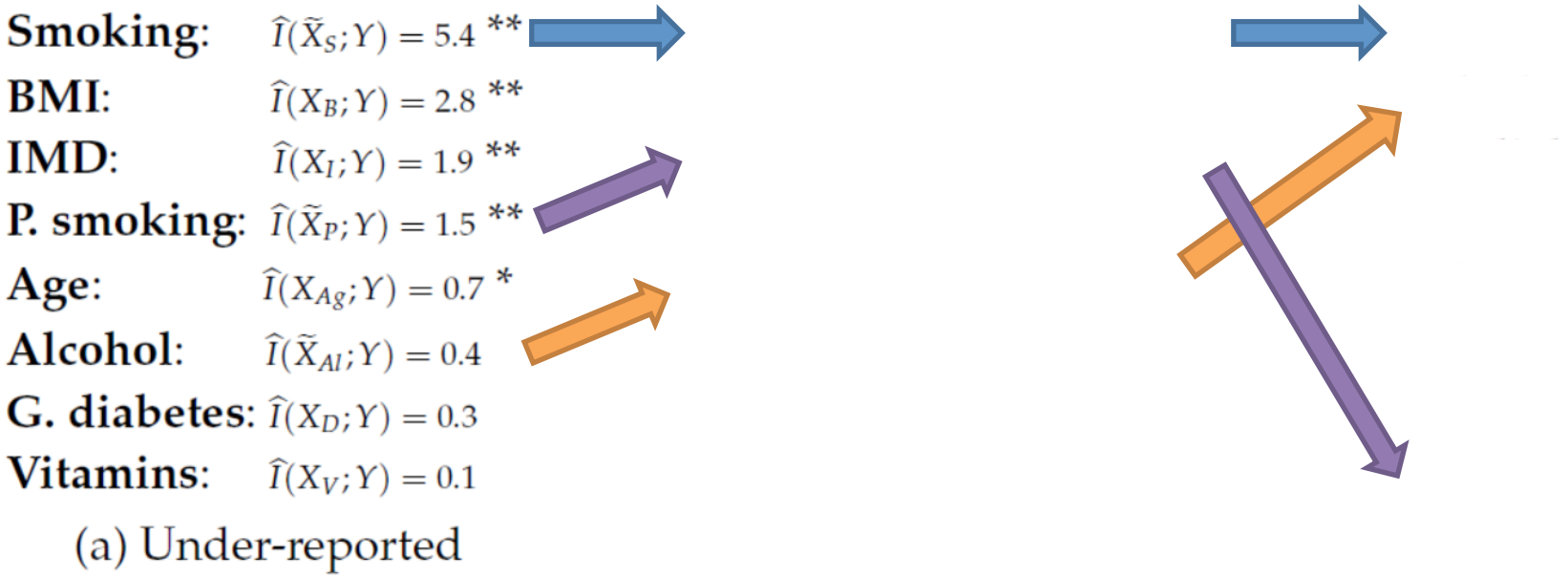


Feature Ranking in UR scenarios



Risk Factors for Low Birth Weight Infants

Risk factors: BMI, IMD, Age, Diabetes, Vitamins, **Smoking, Passive Smoking, Alcohol**



(a) Under-reported

UR are less powerful: Higher Probability of False Negative (Type II error)

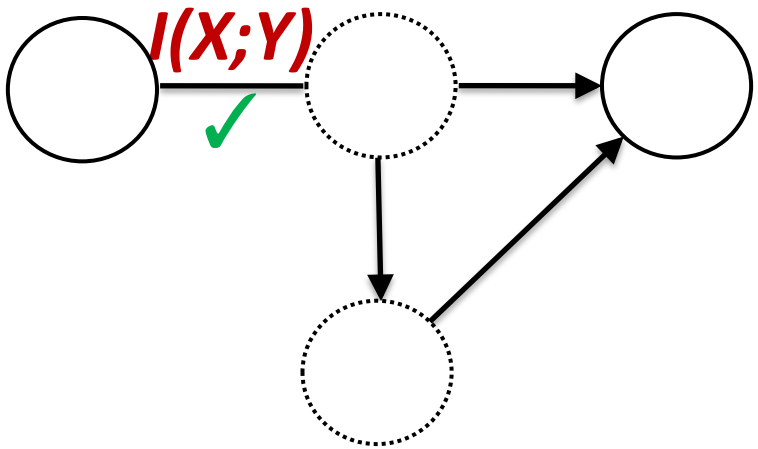
we derived a way to quantify this probability

Ranking that takes into account both Relevancy and Redundancy

mRMR -minimum Redundancy Maximum Relevancy

we derived a way to estimate redundancy between two UR factors

Conclusions and future work



1) Test independence in UR: control False positives/False negatives!

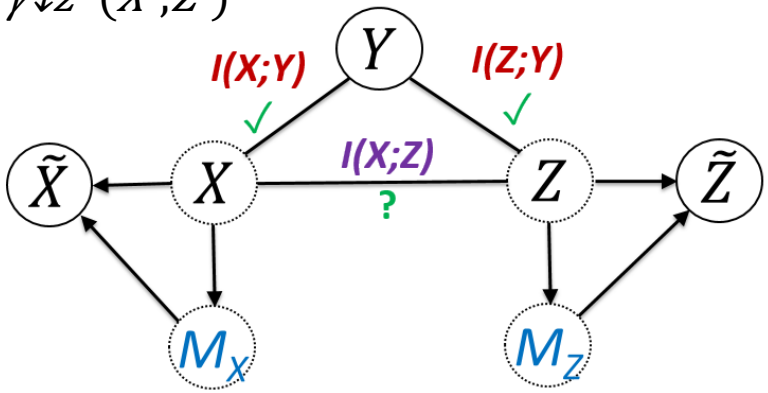
Quantify

effective sample size

2)Estimate redundancy terms: $I(X;Z)=I\downarrow_{\gamma\downarrow x\,\gamma\downarrow z}(X;Z)$

Feature selection

relevancy/redundancy



3) Conditional estimators for MB discovery

Thanks!

Questions?